# Pathway-Structured Predictive Model for Cancer Survival Prediction: A Two-Stage Approach

Xinyan Zhang[1], Yan Li[1], Tomi Akinyemiju[2], Akinyemi I. Ojesina[2], Phillip Buckhaults[3], Nianjun Liu[1], Bo Xu[4], and Nengjun Yi[1, *]

[1] Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL 35294, USA

[2] Department of Epidemiology, University of Alabama at Birmingham, Birmingham, AL 35294, USA

[3] Department of Drug Discovery and Biomedical Sciences, The South Carolina College of Pharmacy, The University of South Carolina, SC 29208, USA

[4] Department of Oncology, Southern Research Institute, Birmingham, Alabama 35205

*** Corresponding author:**

Nengjun Yi

Department of Biostatistics

University of Alabama at Birmingham

Birmingham, AL 35294-0022

Phone: 205-934-4924

Fax: 205-975-2540

Email: nyi@ms.soph.uab.edu

**Key words:** breast cancer prognosis, hierarchical Cox model, pathway, penalized Cox regression, The Cancer Genome Atlas (TCGA)

**Running title:** Pathway-structured prognosis

## Abstract

Heterogeneity in terms of tumor characteristics, prognosis, and survival among cancer patients has been a persistent problem for many decades. Currently, prognosis and outcome predictions are made based on clinical factors and/or by incorporating molecular profiling data. However, inaccurate prognosis and prediction may result by using only clinical or molecular information directly. One of the main shortcomings of past studies is the failure to incorporate prior biological information into the predictive model, given strong evidence of pathway-based genetic nature of cancer, i.e. the potential for oncogenes to be grouped into pathways based on biological functions such as cell survival, proliferation and metastatic dissemination.

To address this problem, we propose a two-stage procedure to incorporate pathway information into the prognostic modeling using large-scale gene expression data. In the first stage, we fit all predictors within each pathway using penalized Cox model (Lasso, Ridge and Elastic Net) and Bayesian hierarchical Cox model. In the second stage, we combine the cross-validated prognostic scores of all pathways obtained in the first stage as new predictors to build an integrated prognostic model for prediction. We apply the proposed method to analyze breast cancer data from The Cancer Genome Atlas (TCGA), predicting overall survival using clinical data and gene expression profiling. The data includes ~20000 genes mapped into 109 pathways for 505 patients. The results show that the proposed approach not only improves survival prediction compared with the alternative analysis that ignores the pathway information, but also identifies significant biological pathways.

## INTRODUCTION

Over the past three decades, remarkable improvement has been achieved in cancer treatment in the United States, with the annual death rate from cancer declining 1.4% for women, 1.8% for men, and 2.3% for children ages 0-10 years from 2002 to 2011 (EDWARDS *et al.* 2014). However, the problem that has persisted in cancer treatment is the heterogeneity of prognostic prediction across patients (BARILLOT 2013). This heterogeneity is, for the most part, genetically determined and rooted in the molecular profile of patients. A Precision medicine initiative has been introduced by White House to expand cancer genomics research as a short-term goal to develop better prevention and treatment methods for more cancers (COLLINS and VARMUS 2015). Recent high-throughput technologies can easily and robustly generated large-scale molecular profiling data, including genomic, epi-genomic, transcriptomic, and proteomic markers offers extraordinary opportunities to integrate clinical and genomic data in prediction models, improving understanding of inter-individual differences that may be critical for the application of precision medicine strategies (BARILLOT 2013; COLLINS and VARMUS 2015).

The recent development of molecular signatures to predict recurrence of breast, colon and prostate cancers are notable and clinically useful but may not be sufficient to achieve the goals of precision medicine (MOOK *et al.* 2007; POHL and LENZ 2008; ENG *et al.* 2013). Gene signatures across different studies with very few overlapped genes can have similar prediction results which suggests there is some underlying mechanism (VAN DE VIJVER *et al.* 2002; WANG *et al.* 2005; SOTIRIOU and PICCART 2007). According to the analysis of 24 pancreatic tumors by Jones et al (2008), altered genes varied greatly across tumors but the pathways with the altered genes remain largely the same, which indicates that the statistical methods focusing on individual genes may be underpowered. It has also been revealed that the genetic nature of cancer is pathway-based, that is, oncogenes can be grouped into pathways based on biological functions such as cell

survival, proliferation and metastatic dissemination (BARILLOT 2013; HUANG *et al.* 2014). Based on the observation that multiple genes in the same biological processes appear to be dysfunctional regardless of cancer type, gene pathways information is likely a more robust biological phenomenon (BILD *et al.* 2006). Various public databases (e.g., KEGG) can be accessed online or in R package to provide the biological information about pathways which may provide valuable improvement to prognosis and prediction (KANEHISA and GOTO 2000). Thus methods incorporating higher-order information of functional units in cancer, i.e., pathways, have been the focus of recent investigations (JONES 2008; JONES *et al.* 2008; LEE *et al.* 2008; REYAL *et al.* 2008; ABRAHAM *et al.* 2010; TESCHENDORFF *et al.* 2010; ENG *et al.* 2013; HUANG *et al.* 2014). Among those previous studies, Abraham et al. adopted a gene set statistic to provide stability of prognostic signatures instead of individual genes (ABRAHAM *et al.* 2010). Huang et al. converted the gene matrix to a pathway matrix through "principal curve", similar to principal components analysis (HUANG *et al.* 2014). Both of these two methods did not incorporate outcome when generating the pathways scores from the individual genes. Other sophisticated statistical methods have been developed for variable selection with grouped predictors or pathways using an "all-in-all-out" idea, meaning that when one predictor in a group is chosen, then all variables in that group are chosen (PARK *et al.* 2007; WEI and LI 2007; JONES 2008). Other methods that could address the above shortcomings have also been developed, however, leading to increased computational complexity and potentially instability of models when the number of predictors is large (HUANG *et al.* 2009; ZHOU 2010; ENG *et al.* 2013). Eng et al. (2013) proposed a method to reduce the computational complexity by incorporating a binary outcome to stand for decreased or increased risk score in each pathway which inferred potentially loss of information.

In this article, to address some of the above shortcomings, we propose a two-stage procedure to incorporate pathway information into the prognostic models using large-scale gene expression data. In the first stage, we fit all predictors within each pathway using penalized Cox model (Lasso, Ridge and Elastic Net) and Bayesian hierarchical Cox model. In the second stage, we combine the cross-validated prognostic scores of all pathways obtained in the first stage as new predictors to build an integrated super prognostic model for prediction. We used the proposed method to analyze a breast cancer data set from The Cancer Genome Atlas (TCGA) project for predicting overall survival using gene expression profiling.

Breast cancer is the second most commonly diagnosed malignancy after skin cancer in women (HUANG *et al.* 2014). It is estimated to be the third leading cause of cancer death after lung cancer and rectal/colon cancer in 2015 (NCI 2015). It is widely understood that breast cancer can be categorized into four clinical subtypes: Luminal A, Luminal B, Triple Negative/Basal like and Her2 and the survival/metastasis outcomes differ significantly among these four subtypes (CAREY *et al.* 2006; O'BRIEN *et al.* 2010; HAQUE *et al.* 2012). However, it is increasingly being realized that using only the clinical subtypes cannot discriminate breast cancers patients, and that better prediction of prognosis is needed. The breast cancer data set from TCGA includes ~20000 genes mapped into 109 pathways for 505 patients. The results show that the proposed approach not only improves survival prediction compared with the alternative analysis that ignores the pathway information, but also identifies significant biological pathways.

## METHODS

### Cox proportional hazards models

Cox regression is the commonly used method for analyzing censored survival data (VAN HOUWELINGGEN and PUTTER 2012), for which the hazard function of survival time $T$ takes the form:

$$h(t \mid X) = h_0(t)\exp(X\beta)$$

where $h_0(t)$ is the baseline hazard function, $X$ and $\beta$ are the vectors of predictors and coefficients, respectively, and $X\beta$ is the linear predictor or called the prognostic index. The coefficients $\beta$ are estimated by maximizing the partial log-likelihood:

$$pl(\beta) = \sum_{i=1}^{n} d_i \log\left(\frac{\exp(X_i\beta)}{\sum_{j\in R(t_i)}\exp(X_j\beta)}\right)$$

where the censoring indicator $d_i$ takes 1 if the observed survival time $t_i$ for individual $i$ is uncensored and 0 if it is censored, and $R(t_i)$ is the risk set at time $t_i$. For molecular data, the number of coefficients is much larger than the number of individuals and/or covariates are usually highly correlated, where Cox regression is not directly applicable.

**Ridge, lasso and elastic-net Cox models.** The elastic net is a widely used penalization approach to handle high-dimensional models, which adds the elastic-net penalty to the log-likelihood function and estimates the parameters $\beta$ by maximizing the penalized log-likelihood(ZOU and HASTIE 2005; HASTIE *et al.* 2009; FRIEDMAN *et al.* 2010; SIMON *et al.* 2011; HASTIE *et al.* 2015). For the Cox models described above, we estimate the parameters $\beta$ by maximizing the penalized partial log-likelihood:

$$ppl_\alpha(\beta) = pl(\beta) - \lambda n\sum_{j=1}^{J}[\alpha \mid \beta_j \mid +(1-\alpha)\frac{1}{2}\beta_j^2]$$

where $\alpha$ ($0 \le \alpha \le 1$) is a predetermined elastic-net parameter, $\lambda$ ($\lambda \ge 0$) is a penalty parameter, and $pl(\beta)$ is the partial log-likelihood of the Cox model. The penalty parameter $\lambda$ controls the overall strength of penalty and the size of the coefficients; for a small $\lambda$, many coefficients can be large, and for a large $\lambda$, many coefficients will be shrunk towards zero. The elastic net includes the lasso ($\alpha = 1$) and ridge Cox regression ($\alpha = 0$) as special cases (TIBSHIRANI 1997; GUI and LI 2005; VAN HOUWELINGEN *et al.* 2006; SIMON *et al.* 2011; VAN HOUWELINGGEN and PUTTER 2012).

The ridge, lasso and elastic net Cox models can be fitted by the cyclic coordinate descent algorithm, which successively optimizes the penalized log-likelihood over each parameter with others fixed and cycles repeatedly until convergence. The cyclic coordinate descent algorithm has been implemented in the R package glmnet. The package glmnet can quickly fit the elastic-net Cox models over a grid of values of $\lambda$ covering the entire range, giving a sequence of models for users to choose from. Cross-validation is the most widely used method to select an optimal value $\lambda$ (e.g., an optimal Cox model) that gives minimum cross-validated error.

**Bayesian hierarchical Cox model.** hierarchical model is an efficient approach to handling high-dimensional data, where the regression coefficients are themselves modeled (GELMAN and HILL 2007; GELMAN *et al.* 2014). Hierarchical models are more easily interpreted and handled in the Bayesian framework where the distribution of the coefficient is the prior distribution, and statistical inference is based on the posterior estimation. The commonly used prior is the double-exponential (or Laplace) prior distribution (PARK and CASELLA 2008; YI and XU 2008; YI and MA 2012):

$$\beta_j \sim DE(\beta_j \mid 0, s) = \frac{1}{2s}\exp\left(-\frac{|\beta_j|}{s}\right)$$

where the scale $s$ is shrinkage parameter and controls the amount of shrinkage; a smaller scale $s$ induces stronger shrinkage and thus forces the estimates of $\beta_j$ towards the prior mean zero. For the hierarchical Cox model with the double-exponential prior, the log posterior distribution of the parameters can be expressed as

$$\log p(\beta \mid t, d) \propto pl(\beta) - \frac{1}{s}\sum_{j=1}^{J}|\beta_j|$$

We fit the hierarchical Cox model by finding the posterior modes of the parameters, i.e., estimating the parameters by maximizing the log posterior distribution. We have developed an algorithm for fitting the hierarchical Cox model by incorporating an EM procedure into the usual Newton-Raphson algorithm for fitting classical Cox models. Our algorithm has been implemented in R package BhGLM (http://www.ssg.uab.edu/bhglm/).

**Relation of hierarchical Cox model to the Lasso**

The Lasso is equivalent to hierarchical Cox model with the double-exponential prior $p(\beta_j) = \frac{n\lambda}{2}\exp(-n\lambda|\beta_j|)$, if one estimates the mode of the posterior distribution. With the double-exponential prior, therefore, the relation between hyper-parameter inverse scale $s$ and penalty tuning parameter $\lambda$ of Lasso is $s = 1/(n\lambda)$.

**Optimizing the penalty by cross-validation and selection of inverse scale**

The penalized Cox model estimate of the previous section depends heavily on the penalty tuning parameter λ. If λ is taken too small, the model boils down to the conventional model virtually with no penalty and the solution still degenerates. If λ is taken too large, all the coefficients are estimated to virtually zero. The tuning parameter is estimated using 10-fold cross-validation over 10 repeats by maximizing the cross-validated partial likelihood (CVPL) $CV(\lambda)$ (VAN HOUWELINGEN *et al.* 2006; SIMON *et al.* 2011):

$$CV(\lambda) = \sum_{k=1}^{K}[pl(\hat{\beta}_{(-k)}) - pl_{(-k)}(\hat{\beta}_{(-k)})]$$

$\hat{\beta}_{(-k)}$ is the estimate of β from all the data except the *k*-th part, $pl(\hat{\beta}_{(-k)})$ is the partial likelihood of all the data points and $pl_{(-k)}(\hat{\beta}_{(-k)})$ is the partial likelihood excluding part *k* of the data. By subtracting the log-partial likelihood evaluated on the non-left out data from that evaluated on the full data, we can make efficient use of the death times of the left out data in relation to the death times of all the data. We choose the λ value which maximizes $CV(\lambda)$. Hyper-parameter inverse scale *s* in hierarchical Cox model is estimated using conversion from estimated tuning parameter λ of Lasso with $s = 1/(n\lambda)$.

**Two-Stage Approach for Pathway Integration**

Intuitively, we can simultaneously fit all the genes in one model. In some previous studies, penalized Cox model (Lasso, Ridge and Elastic Net) have been used to analyze genomic data with this gene-based single model approach (RAPPAPORT 2007; BOVELSTAD *et al.* 2009; JACOB 2009; ZHANG *et al.* 2013; YUAN *et al.* 2014; ZHAO *et al.* 2014). However, due to high-dimension of genomic data, fitting one model including all genes can lead to instability of

predictive model, and may result in bad prediction performance with the increased model complexity.

Here we used an two-stage procedure for combining multiple pathways to build a prediction model, inspired by the super learner of van der Laan et al. (2007) (VAN DER LAAN *et al.* 2007; VAN HOUWELINGEN and PUTTER 2011). The two-stage procedure for building the prognostic model by combining multiple biological pathways is presented in Figure 1.

(Insert Figure 1 here)

Using the notations from Figure 1, gene expression could be divided into $G$ groups (e.g., pathways), $G_g : g = 1, \ldots, G$, with the *g-th* group $G_g$ containing $J_g$ variables. Overlapping is common in this analysis, that is, a gene could belong to multiple functional pathways. In the first stage, we fit all the predictors within g-*th* group (pathway) using the hierarchical Cox model or penalized Cox regression. For *g-th* pathway, we have hazard function: $h(t \mid X_i^g) = h_0(t)\exp(X_i^g \beta^g)$ and obtain the estimate of prognostic score, $\eta_i^g = X_i^g \hat{\beta}^g$, for each individual. To prevent over-fitting, instead of calculating the prognostic indices directly, we estimated cross-validated prognostic scores using leave-one-out cross-validation (LOOCV). The prognostic scores were calculated for *i-th* testing set using the parameters $\hat{\beta}^{g(-i)}$ estimated from (*i*-1) training set. That is, $\eta_{(CV,i)}^g = \eta_i^{g(-i)} = X_i^T \hat{\beta}^{g(-i)}$. In the second stage, we combine the cross-validated prognostic scores from all pathways as new predictors to build a super prognostic model for prediction: $h(t \mid \eta_i) = h_0(t)\exp(\sum_{g=1}^{G} \eta_i^g \alpha^g)$. 10-fold cross-validation over 10 repeats was used to evaluate the predictive performance of the super prognostic model.

**Evaluating the predictive performance**

To assess the prognostic utility of the fitted model, we need to evaluate the quality of the fitted model and its predictive value. There are several ways to measure the performance of a Cox model (STEYERBERG 2009; VAN HOUWELINGGEN and PUTTER 2012): 1) **Concordance Index (C-index)** (HARRELL *et al.* 1996): traditional measurement to determine the concordance between the observed survival times and predicted survival times. The performance is better when the C-index is greater; 2) **CVPL:** as mentioned above as $CV(\lambda)$, is a general measurement of model quality and prediction; 3) **Prediction Error**: Measuring prediction error is an important way to evaluate predictive performance of a survival model. The most popular measure of prediction error is the Brier score, which is defined as (VAN HOUWELINGGEN and PUTTER 2012): $Brier(y, S(t_0 \mid x)) = (y - S(t_0 \mid x))^2$, where $S(t_0 \mid x)$ is the estimated survival probability of an individual beyond $t_0$ given the predictor $x$; 4) **Pre-validated Kaplan Meier Analysis**: We transform the continuous cross-validated prognostic score $\eta_i^{g(-i)}$ from the super prognostic model into a categorical factor based on the median of $\eta_i^{g(-i)}$ and give the Kaplan Meier plot and log-rank test by comparing the 2 groups splitting by median. This allows us to compare statistical significance and prediction performance between models.

## APPLICATION TO TCGA BREAST CANCER DATA

### Data Collection

1071 tumor samples of breast cancer were selected from TCGA breast cancer project. All data including clinical information, microarray mRNA gene expression were downloaded from TCGA as of February 2015 using the *TCGA-Assembler* (ZHU *et al.* 2014). Overall survival (OS) is the outcome of interest. We downloaded and analyzed the processed level 3 (log2 lowess

normalized (cy5/cy3) collapsed by gene symbol) gene expression data. It represents up-regulation or down-regulation of a gene compared relative to the reference population (ZHAO *et al*. 2014).

**Data Preprocessing**

54 samples were removed for missing or zero overall survival time. 17815 features across 533 samples were profiled for gene expression, which includes a total of 1571 missing observations. Simple imputation with mean values across samples was adopted to fill the missing values. For further analysis, only 505 samples were kept for whom survival time and gene expression were both available. Among these 505 patients, only 65 were dead and thus the event rate was 12.9%.

**Pathway Analysis**

To construct the pathways, we used genome annotation tools, KEGG (KANEHISA and GOTO 2000), to map genes to pathways. After mapping gene symbols to Entrez ids, 17252 probes were kept. We mapped all the probes to KEGG pathways using the R annotation package *RDAVIDWebService* (FRESNO and FERNANDEZ 2013). 3181 probes were mapped to 109 pathways.

**Two-Stage Approach for integrating pathways**

**Separately analyzing gene expression data in each pathway.** We compared Lasso, Ridge, Elastic Net ($\alpha = 0.5$) regression and hierarchical Cox models to build predictive model and calculate the LOOCV prognostic scores in each pathway in the first stage. The tuning

parameters in Lasso, Ridge and Elastic Net regressions were estimated by 10-fold cross-validation over 10 repeats. We obtained the estimated tuning parameter λ from Lasso in each pathway. We used $s_1 = 1/n\lambda$, $s_2 = 1/n\lambda + 0.03$ and $s_3 = 0.08$ as the scales of the double-exponential prior for hierarchical Cox model. The table Appendix 1 shows the cross-validated partial likelihood (CVPL) and C-index for each pathway from LOOCV for the hierarchical Cox models ( $s_1 = 1/n\lambda$, $s_2 = 1/n\lambda + 0.03$ and $s_3 = 0.08$, respectively), Lasso, Ridge and Elastic Net Cox regression, respectively.

**Combining all pathways for survival prediction.** In the second stage, only pathways with C-index greater than 0.5 were kept to build a super prognostic model. A super prognostic model was built with the cross-validated prognostic scores of all filtered pathways estimated in the first stage as new predictors. We used hierarchical Cox model with double exponential prior to fit this super prognostic model. 10-fold cross-validation over 10 repeats was also carried out to validate the second stage super prognostic models. To select the prior scales for hierarchical Cox models, we calculated CVPL from 10-fold cross-validation for different prior scales (0.08, 0.10, 0.12, 0.14, 0.16, 0.18, 0.20, and 0.22) and the scale with highest CVPL was chosen. Table 1 shows CVPL, C-index and corresponding prior scales for hierarchical Cox models for all two-stage Lasso-hierarchical Cox Model, two-stage Ridge-hierarchical Cox Model, two-stage Elastic Net-hierarchical Cox Model and three two-stage hierarchical-hierarchical Cox Model ( $s_1 = 1/n\lambda$, $s_2 = 1/n\lambda + 0.03$ and $s_3 = 0.08$, respectively). CVPL for two-stage hierarchical-hierarchical Cox Model ($s_1 = 1/n\lambda$ and $s_2 = 1/n\lambda + 0.03$) and two-stage Lasso-hierarchical Cox Model are -337.170 (7.187), -333.358 (1.971) and -340.058 (4.215). C-index for two-stage hierarchical-hierarchical Cox Model ($s_1 = 1/n\lambda$ and $s_2 = 1/n\lambda + 0.03$) and two-stage Lasso-hierarchical Cox Model are 0.760 (0.015), 0.748 (0.013) and 0.725 (0.014).

(Insert Table 1 here)

**Prediction Performance Comparison between Single Model Analysis and Two-stage Approach**

To compare the predictive performance with the two-stage approach, a joint Lasso and joint hierarchical Cox model were built which simultaneously fit all 3181 genes from 109 pathways. The tuning parameter λ in joint Lasso was estimated to be 0.057 by 10-fold cross-validation over 10 repeats. We used $s = 1/n\lambda$ =0.035 as the scale of the double-exponential prior for joint hierarchical Cox model. 10-fold cross-validation over 10 repeats was used to validate joint Lasso and joint hierarchical Cox model. CVPL and C-index for both joint Lasso and joint hierarchical Cox model are presented in Table 2. CVPL and C-index for joint Lasso are -364.845 (0.949) and 0.507 (0.023), respectively. CVPL and C-index for joint hierarchical Cox model are -363.554 (0.626) and 0.572 (0.023), respectively. Both of them had worse prediction performance than our proposed two-stage combination method.

(Insert Table 2 here)

The predictive performance of the models was also assessed by Brier scores. Figure 1 shows the Brier prediction errors for two-stage hierarchical-hierarchical Cox Model ($s_1 = 1/n\lambda, s_2 = 1/n\lambda + 0.03$), two-stage Lasso-hierarchical Cox Model, and two-stage Ridge-hierarchical Cox Model, the hierarchical Cox fitted pathway with the best predictive performance and joint Lasso. Our two-stage models all had significantly improvement in reducing prediction error compared with best hierarchical Cox fitted pathway or joint Lasso.

(Insert Figure 2 here)

**Pathway Selection**

Figure 3 shows the estimated coefficients and p-values for two-stage Lasso-hierarchical Cox Model and two-stage hierarchical-hierarchical Cox Model ($s_1 = 1/n\lambda$). Pathways with p-values less than 0.05 are labeled in Figure 3. We compared all significant pathways with core cancer pathways discussed in Eng et al. (2013) and listed those consistent core cancer pathways in Table 3.

(Insert Figure 3 and Table 3 here)

**Risk Group Stratification**

In order to demonstrate the potential for using two-stage approach to stratify patients into risk groups, we split the patients by the median of the cross-validated prognostic scores into two groups. Those patients with cross-validated prognostic score greater than the median were categorized as low-risk group; while patients with cross-validated prognostic score less than the median were categorized as high-risk group. The Kaplan-Meier curves for low-risk and high-risk groups from joint Lasso, joint hierarchical Cox model, two-stage Lasso-hierarchical (L-H) model and two-stage hierarchical-hierarchical (H-H) model ($s_1 = 1/n\lambda$) are shown in Figure 4. Log-rank tests are carried out for each model. Both two-stage L-H and H-H models have significant differences of Kaplan Meier curves between low-risk group and high-risk group (p-value = 4.69e-11, p-value = 5.81e-11, respectively). Both joint Lasso and joint hierarchical Cox model have non-significant differences of Kaplan Meier curves between two groups (p-value = 0.242, p-value = 0.231, respectively).

(Insert Figure 4 here)

## DISCUSSION

The heterogeneity of prognostic prediction in cancers has been a persisted problem for decades (BARILLOT 2013). It is now realized that cancer is a fundamentally disease of genome and can be understood by identifying the abnormal genes and proteins that are associated with the risk of developing cancer. Some statistical and machine learning methods have been used to analyze genomic data with gene-based approach to search for gene signature and to predict prognosis (RAPPAPORT 2007; BOVELSTAD *et al.* 2009; JACOB 2009; ZHANG *et al.* 2013; YUAN *et al.* 2014; ZHAO *et al.* 2014). Due to the complicated genetic nature of cancer and potentially underpowered statistical analysis of gene-based approach, it was suggested by Vogelstein that the complexity of cancer should be handled based on pathway-centric instead of gene-centric perspectives (JONES 2008). Oncogenes and tumor suppressor genes have been well studied and can be arranged into signaling pathways according to their biological functions such as cell survival, proliferation and metastatic dissemination. Other studies have investigated methods to analyze high-throughput cancer genomics data based on functional units, i.e. pathways (GOEMAN and BUHLMANN 2007; LEE *et al.* 2008; REYAL *et al.* 2008; ABRAHAM *et al.* 2010; TESCHENDORFF *et al.* 2010).

Our two-stage approach is developed to incorporate the functional structure of pathways to predict survival for cancer patients. Different from some previous methods that summarize pathway score using only gene information or positive/negative signs, our method incorporates the correlation with survival information in the calculating individual pathway information. Besides, we use cross-validated prognostic score as pathway score to be used in the second stage, which not only prevents overfitting and can be easily carried out, but also gives an unbiased view on the contribution of the different information from pathways to the prediction model. Our

approach identified some important pathways, consistent with the core cancer pathways in Eng et al. (2013). Furthermore, all our two-stage approach pathway-based methods performed uniformly better than the gene-based joint models using Lasso or hierarchical cox model in terms of C-index, CVPL and reduced prediction error. Primarily, it can be seen that the highest C-index among the two-stage pathway-based model (0.760) is improved from the performance of both joint gene-based models (C-index = 0.507; 0.572 respectively) by around 0.2. Meanwhile, the Brier prediction errors of all two-stage models have been reduced from gene-based joint models greatly. The Kaplan-Meier analysis between low-risk and high–risk groups dichotomized by two-stage pathway-based models (p-value = 4.69e-11, p-value = 5.81e-11, respectively) have a remarkable performance in discriminating the prognostic effects between different patients compared with both joint gene-based models (p-value = 0.242, p-value = 0.231, respectively). Furthermore, among two-stage methods, the performance is dependent in the method selected to calculate the pathway score in the first stage. Lasso and hierarchical Cox models with the two estimated priors ($s_1 = 1/n\lambda$ and $s_2 = 1/n\lambda + 0.03$) perform uniformly better than Ridge, Elastic Net and hierarchical Cox model with prior ($s_3 = 0.08$).

Our approach is also capable of identifying core pathways in cancer. Mitogen-activated protein kinase (MAPK) pathways are important in controlling fundamental cellular processes, i.e. growth, proliferation, differentiation, migration and apoptosis (DHILLON *et al.* 2007). When abnormally activated, MAPK pathways can lead to the progression of cancer (USSAR and VOSS 2004; MCCUBREY *et al.* 2007). Another pathway, the mammalian target of rapamycin (mTOR), also plays an essential role in the regulation of cell proliferation, growth, differentiation, migration and survival. Similarly to MAPK pathways, the dysregulation of mTOR signaling happens in various human tumors, resulting in higher susceptibility to inhibitors of mTOR

(HUANG and HOUGHTON 2003). The Hedgehog pathway regulates many fundamental processes including stem cell maintenance, cell differentiation, tissue polarity and cell proliferation. It has been demonstrated that inappropriate activation of Hedgehog pathway occurs in various cancers such as brain, gastrointestinal, lung, breast and prostate cancers (GUPTA *et al.* 2010). Furthermore, the JAK-STAT pathway is also identified in our approach. This pathway regulates in various cellular processes such as stem cell maintenance, apoptosis and the inflammatory response and was found frequently dysregulated in diverse types of cancer (THOMAS *et al.* 2015).

However, there are some potential limitations in this method. We implemented our approach only in microarray gene expression data from TCGA breast cancer project. There are many other platforms in gene expression data, such as RNA-Seq data. It may require additional constraints for the RNA-Seq data to be implemented into the model. Another potential limitation is that only 3181 mapped genes among the nearly 20000 expressed genes were fitted in the model, suggesting potentially loss of gene information.

Despite the above potential limitations, our two-stage pathway-based approach performs better in predicting overall survival of breast cancer and is able to identify important cancer pathways. In the future, we will apply our approach in other levels of genomic data, e.g. DNA methylation, miRNA and copy number alterations, for more than 30 types of cancer. We will also continue to develop more efficient ways to combine different levels of genomic data as well as clinical biomarkers into our proposed method to better predict cancer survival.

## Acknowledgments

This work was supported in part by the research grants: NIH 2 R01GM069430.

## REFERENCES


Abraham, G., A. Kowalczyk, S. Loi, I. Haviv and J. Zobel, 2010 Prediction of breast cancer prognosis using gene set statistics provides signature stability and biological context. BMC Bioinformatics 11: 277.

Barillot, E., 2013 *Computational systems biology of cancer*. CRC Press, Boca Raton, FL.

Bild, A. H., G. Yao, J. T. Chang, Q. Wang, A. Potti *et al.*, 2006 Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature 439: 353-357.

Bovelstad, H. M., S. Nygard and O. Borgan, 2009 Survival prediction from clinico-genomic models--a comparative study. BMC Bioinformatics 10: 413.

Carey, L. A., C. M. Perou, C. A. Livasy, L. G. Dressler, D. Cowan *et al.*, 2006 Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. JAMA 295: 2492-2502.

Collins, F. S., and H. Varmus, 2015 A new initiative on precision medicine. N Engl J Med 372: 793-795.

Dhillon, A. S., S. Hagan, O. Rath and W. Kolch, 2007 MAP kinase signalling pathways in cancer. Oncogene 26: 3279-3290.

Edwards, B. K., A. M. Noone, A. B. Mariotto, E. P. Simard, F. P. Boscoe *et al.*, 2014 Annual Report to the Nation on the status of cancer, 1975-2010, featuring prevalence of comorbidity and impact on survival among persons with lung, colorectal, breast, or prostate cancer. Cancer 120: 1290-1314.

Eng, K. H., S. Wang, W. H. Bradley, J. S. Rader and C. Kendziorski, 2013 Pathway index models for construction of patient-specific risk profiles. Stat Med 32: 1524-1535.

Fresno, C., and E. A. Fernandez, 2013 RDAVIDWebService: a versatile R interface to DAVID. Bioinformatics 29: 2810-2811.

Friedman, J., T. Hastie and R. Tibshirani, 2010 Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 33: 1-22.

Gelman, A., J. B. Carlin, H. S. Stern and D. B. Rubin, 2014 *Bayesian data analysis*. Taylor & Francis.

Gelman, A., and J. Hill, 2007 Data analysis using regression and hierarchical/multilevel models, pp. Cambridge University Press: Cambridge, UK.

Goeman, J. J., and P. Buhlmann, 2007 Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics 23: 980-987.

Gui, J., and H. Li, 2005 Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. Bioinformatics 21: 3001-3008.

Gupta, S., N. Takebe and P. Lorusso, 2010 Targeting the Hedgehog pathway in cancer. Ther Adv Med Oncol 2: 237-250.

Haque, R., S. A. Ahmed, G. Inzhakova, J. Shi, C. Avila *et al.*, 2012 Impact of breast cancer subtypes and treatment on survival: an analysis spanning two decades. Cancer Epidemiol Biomarkers Prev 21: 1848-1855.

Harrell, F. E., Jr., K. L. Lee and D. B. Mark, 1996 Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 15: 361-387.

Hastie, T., R. Tibshirani and J. Friedman, 2009 *The Elements of Statistical Learning*. Springer-Verlag, New York, NY, USA.

Hastie, T., R. Tibshirani and M. Wainwright, 2015 *Statistical Learning with Sparsity - The Lasso and Generalization*. CRC Press, New York.

Huang, J., S. Ma, H. Xie and C. H. Zhang, 2009 A group bridge approach for variable selection. Biometrika 96: 339-355.

Huang, S., and P. J. Houghton, 2003 Targeting mTOR signaling for cancer therapy. Curr Opin Pharmacol 3: 371-377.

Huang, S., C. Yee, T. Ching, H. Yu and L. X. Garmire, 2014 A novel model to combine clinical and pathway-based transcriptomic information for the prognosis prediction of breast cancer. PLoS Comput Biol 10: e1003851.

Jacob, L. e. a., 2009 Group lasso with overlap and graph lasso.

Jones, D., 2008 Pathways to cancer therapy. Nat Rev Drug Discov 7: 875-876.

Jones, S., X. Zhang, D. W. Parsons, J. C. Lin, R. J. Leary *et al.*, 2008 Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. Science 321: 1801-1806.

Kanehisa, M., and S. Goto, 2000 KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28: 27-30.

Lee, E., H. Y. Chuang, J. W. Kim, T. Ideker and D. Lee, 2008 Inferring pathway activity toward precise disease classification. PLoS Comput Biol 4: e1000217.

McCubrey, J. A., L. S. Steelman, W. H. Chappell, S. L. Abrams, E. W. Wong *et al.*, 2007 Roles of the Raf/MEK/ERK pathway in cell growth, malignant transformation and drug resistance. Biochim Biophys Acta 1773: 1263-1284.

Mook, S., L. J. Van't Veer, E. J. Rutgers, M. J. Piccart-Gebhart and F. Cardoso, 2007 Individualization of therapy using Mammaprint: from development to the MINDACT Trial. Cancer Genomics Proteomics 4: 147-155.

NCI, 2015 Breast Cance – for patients, pp. National Cancer Institute.

O'Brien, K. M., S. R. Cole, C. K. Tse, C. M. Perou, L. A. Carey *et al.*, 2010 Intrinsic breast tumor subtypes, race, and long-term survival in the Carolina Breast Cancer Study. Clin Cancer Res 16: 6100-6110.

Park, M. Y., T. Hastie and R. Tibshirani, 2007 Averaged gene expressions for regression. Biostatistics 8: 212-227.

Park, T., and G. Casella, 2008 The Bayesian Lasso. Journal of the American Statistical Association 103: 681-686.

Pohl, A., and H. J. Lenz, 2008 Individualization of therapy for colorectal cancer based on clinical and molecular parameters. Gastrointest Cancer Res 2: S38-41.

Rappaport, F. e. a., 2007 Classification of microarray data using gene networks. BMC Bioinformatics 8.

Reyal, F., M. H. van Vliet, N. J. Armstrong, H. M. Horlings, K. E. de Visser *et al.*, 2008 A comprehensive analysis of prognostic signatures reveals the high predictive capacity of the proliferation, immune response and RNA splicing modules in breast cancer. Breast Cancer Res 10: R93.

Simon, N., J. Friedman, T. Hastie and R. Tibshirani, 2011 Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. Journal of Statistical Software 39: 1-13.

Sotiriou, C., and M. J. Piccart, 2007 Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care? Nat Rev Cancer 7: 545-553.

Steyerberg, E. W., 2009 *Clinical Prediction Models: A Practical Approch to Development, Validation, and Updates.* Springer, New York.

Teschendorff, A. E., S. Gomez, A. Arenas, D. El-Ashry, M. Schmidt *et al.*, 2010 Improved prognostic classification of breast cancer defined by antagonistic activation patterns of immune response pathway modules. BMC Cancer 10: 604.

Thomas, S. J., J. A. Snowden, M. P. Zeidler and S. J. Danson, 2015 The role of JAK/STAT signalling in the pathogenesis, prognosis and treatment of solid tumours. Br J Cancer 113: 365-371.

Tibshirani, R., 1997 The lasso method for variable selection in the Cox model. Stat Med 16: 385-395.

Ussar, S., and T. Voss, 2004 MEK1 and MEK2, different regulators of the G1/S transition. J Biol Chem 279: 43861-43869.

van de Vijver, M. J., Y. D. He, L. J. van't Veer, H. Dai, A. A. Hart *et al.*, 2002 A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 347: 1999-2009.

van der Laan, M. J., E. C. Polley and A. E. Hubbard, 2007 Super learner. Statistical applications in genetics and molecular biology 6.

van Houwelingen, H., and H. Putter, 2011 *Dynamic prediction in clinical survival analysis*. CRC Press.

van Houwelingen, H. C., T. Bruinsma, A. A. Hart, L. J. Van't Veer and L. F. Wessels, 2006 Cross-validated Cox regression on microarray gene expression data. Stat Med 25: 3201-3216.

van Houwelinggen, H. G., and H. Putter, 2012 *Dynamic Prediction in Clinical Survival Analysis*. CRC Press.

Wang, Y., J. G. Klijn, Y. Zhang, A. M. Sieuwerts, M. P. Look *et al.*, 2005 Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 365: 671-679.

Wei, Z., and H. Li, 2007 Nonparametric pathway-based regression models for analysis of genomic data. Biostatistics 8: 265-284.

Yi, N., and S. Ma, 2012 Hierarchical Shrinkage Priors and Model Fitting for High-dimensional Generalized Linear Models. Stat Appl Genet Mol Biol.

Yi, N., and S. Xu, 2008 Bayesian LASSO for quantitative trait loci mapping. Genetics 179: 1045-1055.

Yuan, Y., E. M. Van Allen, L. Omberg, N. Wagle, A. Amin-Mansour *et al.*, 2014 Assessing the clinical utility of cancer genomic and proteomic data across tumor types. Nat Biotechnol 32: 644-652.

Zhang, W., T. Ota, V. Shridhar, J. Chien, B. Wu *et al.*, 2013 Network-based survival analysis reveals subnetwork signatures for predicting outcomes of ovarian cancer treatment. PLoS Comput Biol 9: e1002975.

Zhao, Q., X. Shi, Y. Xie, J. Huang, B. Shia *et al.*, 2014 Combining multidimensional genomic measurements for predicting cancer prognosis: observations from TCGA. Brief Bioinform 16: 291-303.

Zhou, N. Z., J., 2010 Group Variable Selection via a Hierarchical Lasso and Its Oracle Property, pp. Arxiv preprint.

Zhu, Y., P. Qiu and Y. Ji, 2014 TCGA-assembler: open-source software for retrieving and processing TCGA data. Nat Methods 11: 599-600.

Zou, H., and T. Hastie, 2005 Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B 67: 301–320.

## Figure Legends

**Figure 1.** Flowchart of the two-stage prognostic model.

**Figure 2.** Brier prediction errors for Two-stage hierarchical-hierarchical Cox Model ($s_1 = 1/n\lambda$ and $s_2 = 1/n\lambda + 0.03$ ), Two-stage Lasso-hierarchical Cox Model, Two-stage Ridge-hierarchical Cox Model, Best hierarchical Cox Fitted Pathway, and joint Lasso, listed successively as black, blue, red, green, light blue, and yellow dash lines.

**Figure 3.** Estimated Coefficients and P-values of two-stage Lasso-hierarchical Cox model and two-stage hierarchical-hierarchical Cox model ($s_1 = 1/n\lambda$), left column is the list of pathways; while the right column is the p-value for each pathway.

**Figure 4.** Kaplan-Meier Curves for low-risk and high-risk groups from joint Lasso, joint hierarchical Cox model, two-stage Lasso-hierarchical (L-H) model and two-stage hierarchical-hierarchical (H-H) model ($s_1 = 1/n\lambda$), p-values are calculated using log-rank test. Red dash line is for all tumors; green solid line is for low-risk group; blue solid line is for high-risk group.

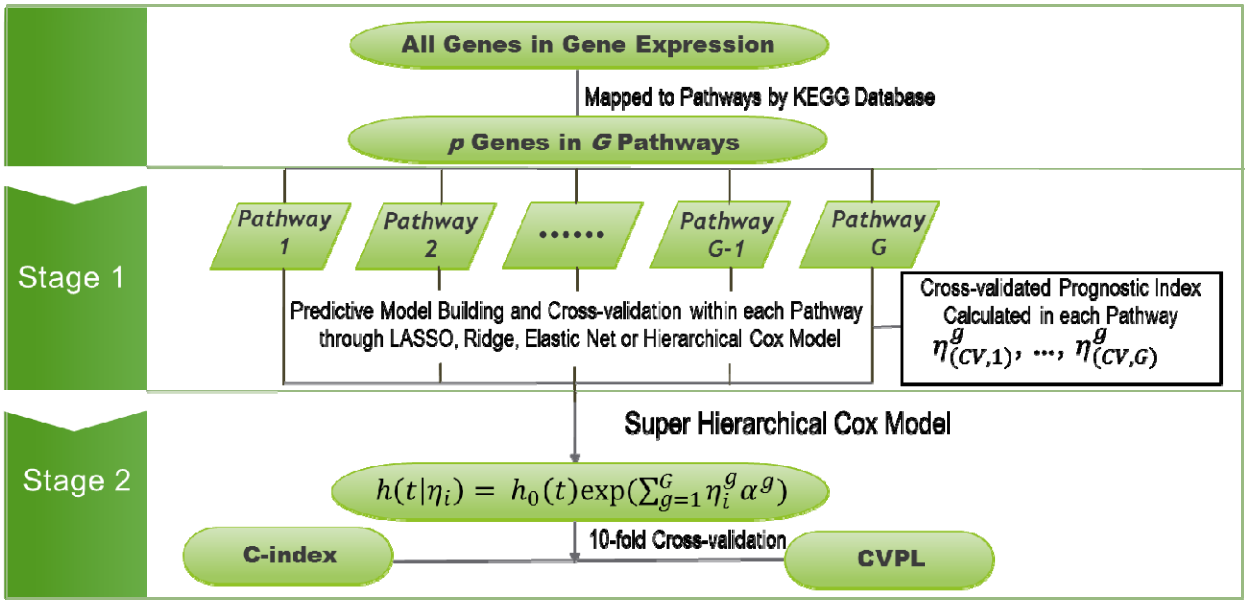


**Figure 1. Flowchart of the two-stage prognostic model.**

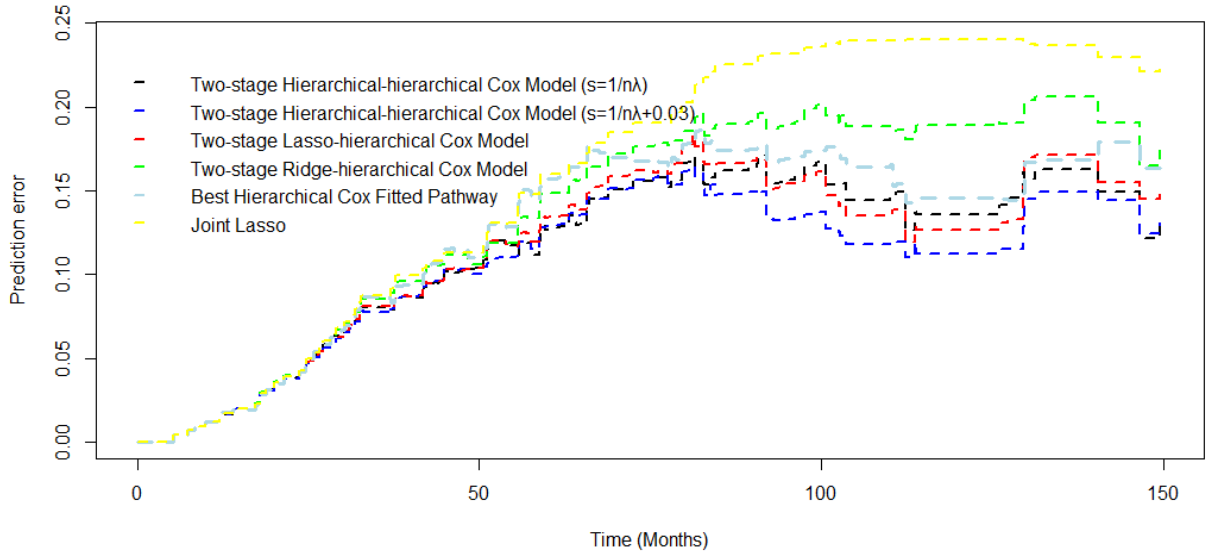


**Figure 2.** Brier prediction errors for Two-stage hierarchical-hierarchical Cox Model ($s_1 = 1/n\lambda$ and $s_2 = 1/n\lambda + 0.03$ ), Two-stage Lasso-hierarchical Cox Model, Two-stage Ridge-hierarchical Cox Model, Best hierarchical Cox Fitted Pathway, and joint Lasso.

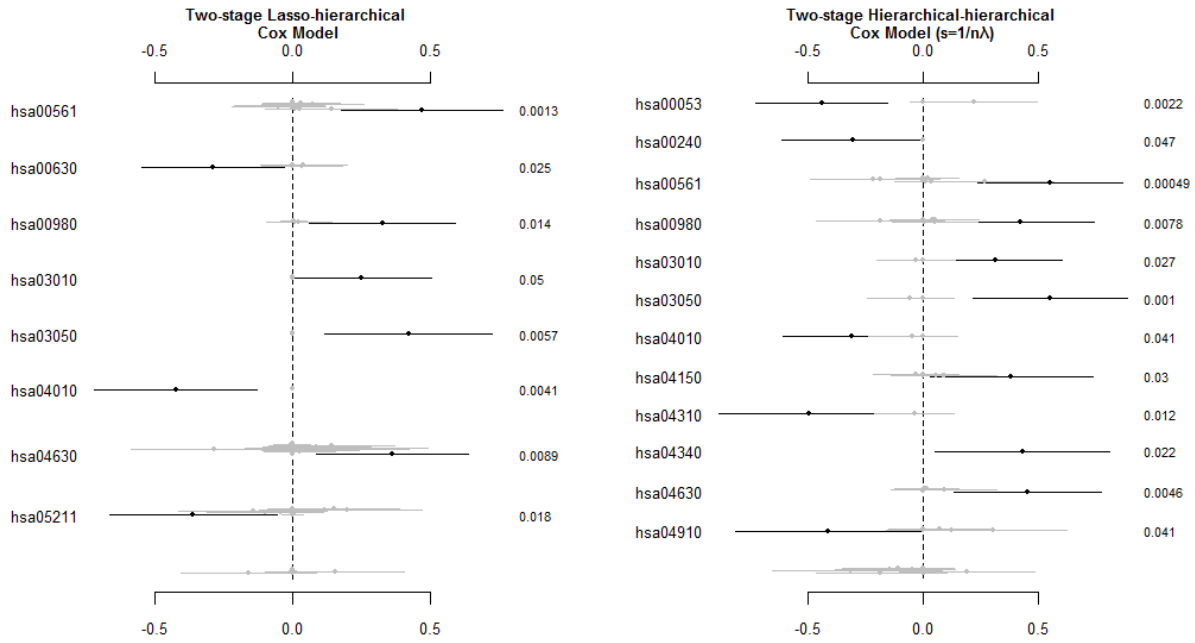


**Figure 3.** Estimated Coefficients and P-values of two-stage Lasso-hierarchical Cox model and two-stage hierarchical-hierarchical Cox model ($s_1 = 1/n\lambda$).

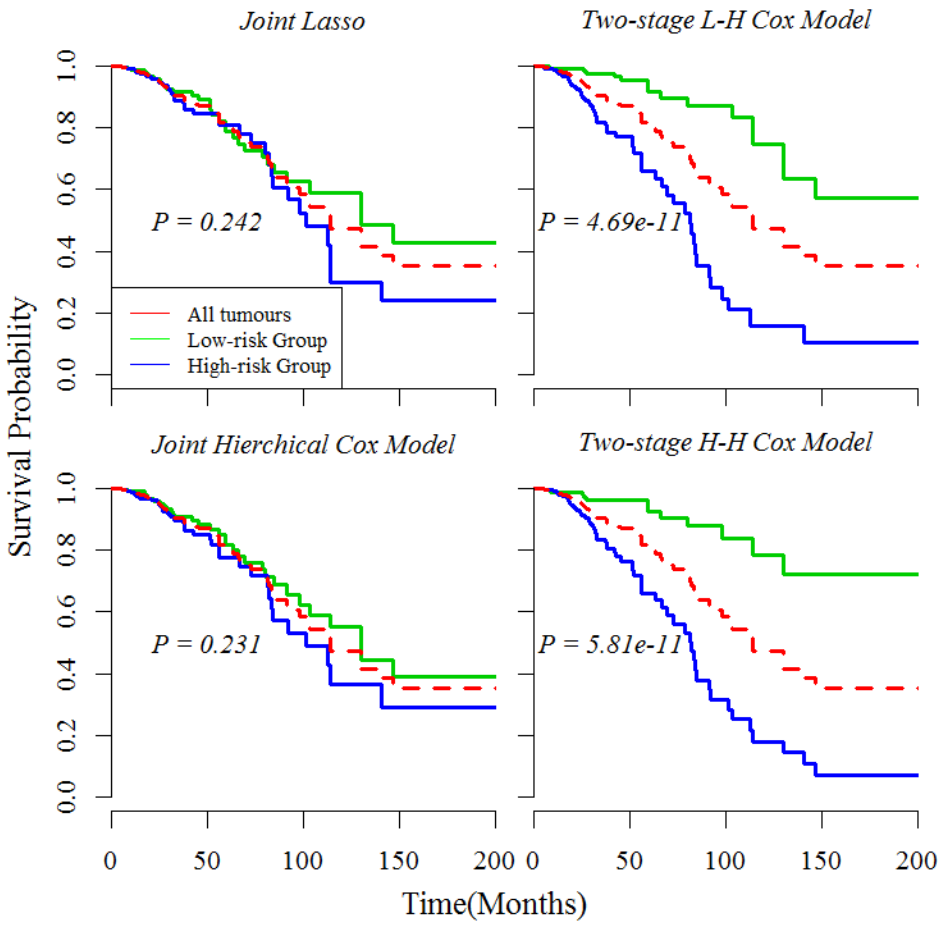


**Figure 4.** Kaplan-Meier Curves for low-risk and high-risk groups from joint Lasso, joint hierarchical Cox model, two-stage Lasso-hierarchical (L-H) model and two-stage hierarchical-hierarchical (H-H) model ($s_1 = 1/n\lambda$), p-values are calculated using log-rank test.

**Table 1. Prediction Performance Comparison between Different Two-stage Models.**

| Model | Number of Pathways | CVPL | C-index | Prior Scale Second Stage hierarchical Model |
|---|---|---|---|---|
| Two-stage Lasso-hierarchical Cox Model | 75 | -340.058 (4.215) | 0.725 (0.014) | 0.20 |
| Two-stage Ridge-hierarchical Cox Model | 88 | -353.444 (2.373) | 0.640 (0.021) | 0.10 |
| Two-stage Elastic Net-hierarchical Cox Model ( $\alpha$=0.5) | 74 | -340.893 (2.435) | 0.711 (0.012) | 0.16 |
| Two-stage hierarchical-hierarchical Cox Model ($s_1 = 1/n\lambda$ ) | 77 | -337.170 (7.187) | 0.760 (0.015) | 0.26 |
| Two-stage hierarchical-hierarchical Cox Model ($s_2 = 1/n\lambda + 0.03$) | 77 | -333.358 (1.971) | 0.748 (0.013) | 0.14 |
| Two-stage hierarchical-hierarchical Cox Model ($s_3 = 0.08$) | 70 | -347.459 (2.796) | 0.692 (0.014) | 0.16 |

**Table 2. Prediction performance comparison between two single models.**

| Model | Joint Lasso | Joint Hierarchical Cox Model |
|---|---|---|
| Number of Genes* | 3181 | 3181 |
| CVPL | -364.845 (0.949) | -363.554 (0.626) |
| C-index | 0.507 (0.023) | 0.572 (0.023) |

**Table 3. Pathway selection comparison between two-stage Lasso-hierarchical Cox model and two-stage hierarchical-hierarchical Cox model ($s_1 = 1/n\lambda$).**

| Two-stage Lasso-hierarchical model | Two-stage Hierarchical-hierarchical model ($s_1 = 1/n\lambda$) |
|---|---|
| hsa04010:MAPK signaling pathway | hsa04010:MAPK signaling pathway |
| hsa04630:Jak-STAT signaling pathway | hsa04150:mTOR signaling pathway |
| | hsa04340:Hedgehog signaling pathway |
| | hsa04630:Jak-STAT signaling pathway |